NAMED ENTITY EXTRACTION FROM SPEECH: APPROACH AND RESULTS USING THE TEXTPRO SYSTEM

Douglas E. Appelt
David Martin

Artificial Intelligence Center
SRI International
333 Ravenswood Ave,
Menlo Park, CA 94025

ABSTRACT

This paper describes the application of the TextPro system to the task of recognition of named entities in speech. TextPro is a lightweight engine for interpreting cascaded finite-state transducers. Although originally intended for processing text, the experience of this evaluation demonstrates the system can easily be adapted to processing transcripts generated by a speech recognizer as well.

1. THE TEXTPRO EXTRACTION SYSTEM

For its participation in the Hub4 named-entity identification task, SRI International employed a newly developed information extraction system called TextPro. TextPro is a lightweight interpreter of cascaded finite-state transducers that is based on the TIPSTER Document Manager architecture [Grishman et al., 1996] and the TIPSTER Common Pattern Specification Language[1] (CPSL).

TextPro finite-state transducers accept and produce sequences of annotations on the document conforming to the structure specified by the TIPSTER document manager architecture. The transducers themselves are expressed by finite-state rules written in CPSL. The grammars employed by the Hub4 name recognizer specified the creation of ENAMEX, NUMEX, and TIMEX annotations, as well as other annotations used by the system internally. After having run each of the cascaded transducers over an input text, a postprocessor would insert SGML markup as required by the rules of the named-entity task.

TextPro was originally developed to process text documents, and to test alternative specifications for CPSL. The first author participated in the design committee for CPSL under the TIPSTER program. The program runs on PowerPC Macintosh computers, and is freely downloadable from the World Wide Web.[2] Although originally developed for limited objectives, experience led us to conclude that TextPro was a very useful system for performing document annotation tasks that do not involve the construction and merging of template structures such as those typical of the MUC scenario template tasks [ref Muc6]. For this reason, and because it is small and extremely fast, we felt that TextPro was a superior alternative to the more well known FASTUS system [Hobbs et al., 1996], which SRI has employed in various MUC evaluations.

1.1. Adapting TextPro to the Hub4 task

Although TextPro was originally intended to process newspaper texts, it proved to be very straightforward to process speech transcriptions in Universal Transcription Format, whether human or machine generated. The adaptation process began by translating the FASTUS grammar used for SRI International's participation in MUC-6 into CPSL. The MUC-6 grammar provided a high-performance baseline to start from; the SRI MUC-6 FASTUS system performed well on the named-entity task, achieving an F-measure of 94. The MUC-6 name recognizer, however, was optimized for mixed case texts, and typical Wall Street Journal articles, and therefore its performance on the Hub4 task was considerably short of optimal.

Adapting the grammar to work well with monocase texts, absent any information provided by capitalization in ordinary texts, required the use of large lexicons to indicate which words were likely to be parts of names. The TextPro Hub4 system uses four large lexicons in addition to the lexicons used by the MUC-6 system:

1. A large lexicon of United States place names that was originally distributed with the place-name gazetteer for MUC-5, supplemented by a manually-culled set of foreign place names from the same source.

2. A proprietary list of person names of many nationalities obtained from Nuance Communications Corp.[3]

[1] Because of the premature end of the TIPSTER program, the specifications for the Common Pattern Specification Language were never finalized or published. Further information is obtainable from the authors.

[2] The URL for obtaining TextPro is http://www.ai.sri.com/~appelt/TextPro/.

[3] This proprietary list cannot be given out in a public distribution. It is possible to replace this list with a list of American first and last names culled from census data publicly available on the Web. However, because of the

3. A list of prominent American and multinational corporations. This lexicon was the same one used by SRI for MUC-5 and MUC-6 participation, with the addition of some recently founded corporations.

4. A list of United States government agencies and departments. The lexicon used in this evaluation was expanded considerably over previous versions by using names appearing in the Hub4 training data.

After the initial grammars and large lexicons were in place, the next step was to do iterative testing and debugging to raise the level of the system's performance by refining the rules and lexical entries. Given the high speed of the TextPro system, it was very easy to do runs over the entire set of training data. Ten megabytes of training data could be processed in about two hours on our available hardware.

The process of hill climbing on training data was not much different for the Hub4 task than other information extraction tasks in which SRI has participated. A few innovations were necessary to process speech transcription data successfully:

- Important discourse contexts, in particular sports report and weather report contexts, were recognized. Sports report contexts were recognized so that names referring to sports teams that were ambiguous with ordinary English words (e.g. "Indians") would be properly identified. In weather report contexts, it was important to recognize that phrases like "in the sixties" are not temporal expressions.

- Rules were needed to decompose lists of person-name words into likely combinations of first, middle, and last names. This was very important for lists or conjunctions of person names, and for the frequent situations where names appeared adjacent to a sentence boundary that would not be marked in the speech transcript.

- Successful name recognition requires recognizing subsequent references to the same person, particularly when such references involve only the first name or the last name of a previously mentioned person that would not otherwise be tagged as names because of ambiguity with ordinary English words. This strategy works very well, except in the relatively frequent situations in which the speaker utters a fragment, or a repair. These fragments, if incorrectly recognized as names, can cause the erroneous tagging of many words in a text when the fragments are recognized as common one-syllable words. The TextPro Hub4 system used frequency data gleaned from the Penn Treebank Wall Street Journal corpus to limit the recognition of very common words as name parts to those contexts in which their status as names was unambiguous.

1.2. Evaluation Results

The table in Figure 1 illustrates the results obtained by the TextPro Hub4 system in the recent evaluation for TextPro applied to the reference transcripts, and for TextPro applied to the output of SRI's own speech recognition system. For the reference transcripts, and the baseline recognizer output, the TextPro results are very close to the best reported in each category.

These evaluation results are quite consistent with the results obtained by SRI during our development testing. For development and testing, we divided the available 10 megabytes of training data furnished by Mitre and BBN into an eight-

Evaluation Task	Content	Extent	Type	Average
Reference transcript, Segment 1	0.93	0.87	0.90	0.90
Reference transcript, Segment 2	0.93	0.88	0.91	0.91
SRI recognizer, Segment 1	0.76	0.75	0.79	0.77
SRI recognizer, Segment 2	0.80	0.76	0.81	0.79
SRI <10X Real Time Recognizer	0.76	0.74	0.78	0.76

Figure 1: SRI International's TextPro tagging results (F-measures) on the Hub4 named entity recognition task

census data's weaker coverage of foreign names, its performance on the Hub4 task is noticeably lower.

Source	Percentage of Errors	Remarks
Case	38.6%	These errors are the direct result of the ambiguity that arises in handling monocase text.
Undercoverage	31.3%	These are errors that could probably be fixed by the addition of a new rule or a feature on a lexical item.
Genre	10.5%	These errors are introduced by the ambiguity that arises in a less formal genre than newspaper text, typical of spoken language in the broadcast news domain.
Evaluation	9.4%	These errors are due to elements of the named entity extraction specification that were intentionally ignored during development, or disagreement with the annotator's interpretation of the rules.
Speech	8.4%	These errors were introduced by using spoken language as the input to the system. It includes word fragments, false starts, and lack of sentence-final punctuation.
Other	1.8%	Errors that don't fit into any of the above categories.

Figure 2: Coarse analysis of sources of TextPro named entity recognition errors

megabyte training corpus and a two-megabyte test corpus, which was kept blind. In a final run before the official test, we recorded an average F-measure of 92 for the training data, and 89 for the blind development test data.

2. ERROR ANALYSIS

It is illuminating to examine a representative cross section of the development test data to discover precisely what the sources of error were in processing the reference transcripts. For the sake of this analysis, one can assume that the word error rate for human-generated transcripts is essentially zero.

Figure 2 divides the sources of error into six different categories. This division recognizes the mixed-case newspaper text for which TextPro was originally developed as a "baseline." The Hub4 task is inherently more difficult because it introduces sources of error that are not present in newspaper text, coming from monocase text, a less formal genre, word fragments, and lack of sentence endpoints that characterize the speech transcription data.

The percentages in the table indicate the *percentage of errors* that are accounted for by each error source.

DISCUSSION OF RESULTS

We believe that these results demonstrate that named entity recognition from speech is not likely to be significantly improved from the levels of performance achieved in this evaluation solely by improvements to the named-entity recognition modules alone. Perhaps tighter integration between speech recognition and a named-entity recognizer could improve the accuracy of the speech recognizer. We treated the two processes as independent modules for this evaluation.

3.1 Limited Possibilities for Improvement

We believe that some of the categories of error indicated in Figure 2 represent a nearly irreducible source of error. The "Evaluation" error source is accounted for by a level of interannotator disagreement that is consistent with the disagreement that has been measured for similar tasks in MUC. It is quite easy to fix the errors in the "undercoverage" category, but it is far from clear that overall performance on novel input would be improved. Experience from the later stages of development for this evaluation shows that the tail of the distribution of new lexical items and patterns is long enough that continued work in this area is not reflected in significantly better results. The ambiguities introduced by monocase text and informal genre generally present difficult contextual resolution problems that are resistant to simple solutions. The problems introduced by processing the output of a speech recognizer may well be solvable with a tighter integration between the speech recognition and name tagging modules. However, the relatively small number of errors that are traceable to this source limits the upside potential of this integration.

3.2 Advantages of Rule-Based Name Recognition

This evaluation provides a good comparison between rule-based and statistical approaches to name recognition, and it suggests that both approaches top out at about the same level of performance. The name recognition task is an ideal application for statistical techniques because training data is available in large quantities, and if desired, more can be obtained with relatively little effort. Large lexicons of proper names are available from public-domain sources, at least for English.

However, our experience suggests that rule based systems are suitable for certain kinds of name recognition tasks, particularly

ones where small changes in the specifications for the information being sought must be accommodated. Certain kinds of specification changes can be accommodated quickly and easily with rule modification. For example, if one decides that astronomical objects really shouldn't be considered locations on a par with "California," then this change can be implemented my modifying a single rule in TextPro. It is less clear how such a change in specifications would be accommodated in an automatically trained system, but in the worst case, it could involve reviewing and reannotating a large quantity of data.

On the other hand, certain changes in specifications can be quite difficult for rule-based systems. For example, switching from mixed case to monocase text can require rewriting a substantial number of rules, and it can be easily argued that retraining is much easier.

Ultimately, the requirements of a particular application will determine the approach to be used. It is reassuring to have some evidence that similar results can be obtained.

REFERENCES

1. Grishman, R. et al., *TIPSTER Text Phase II Architecture Design*, September, 1996.
2. Hobbs et al., *FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text*, in Roche and Schabes (eds.) *Finite State Devices for Natural Language Processing*, MIT Press, 1996.